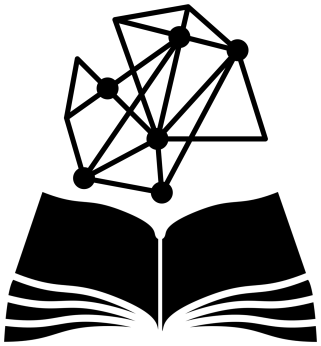
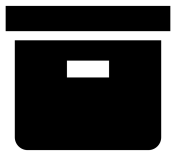


Visualizing Library Resources as Networks

Matt Miller
matthewmiller@nypl.org | @thisismmiller
NYPL Labs
New York Public Library

Agenda

Why Networks?

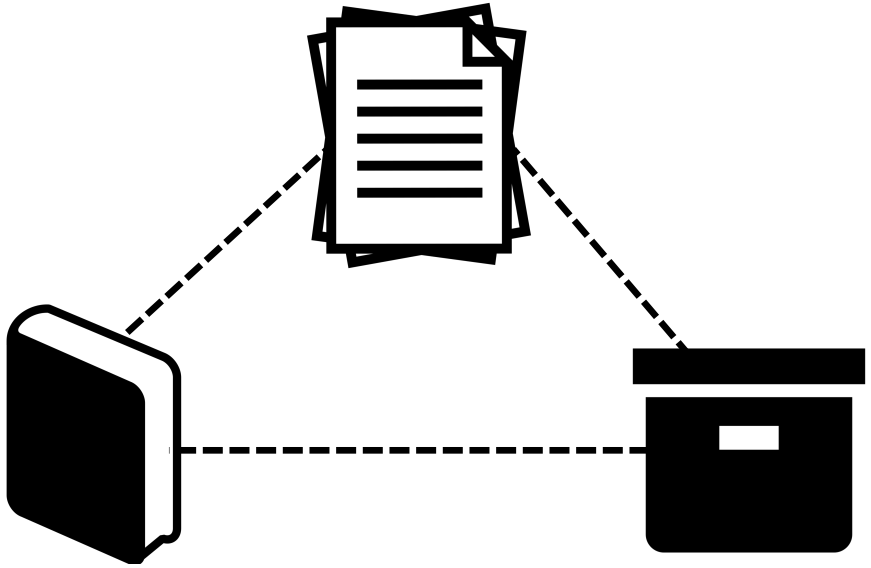


Networks in Archives

Networks in the Catalog

Why Networks?

- Libraries and archives consist of independent documents. How do we connect them and why is it important?

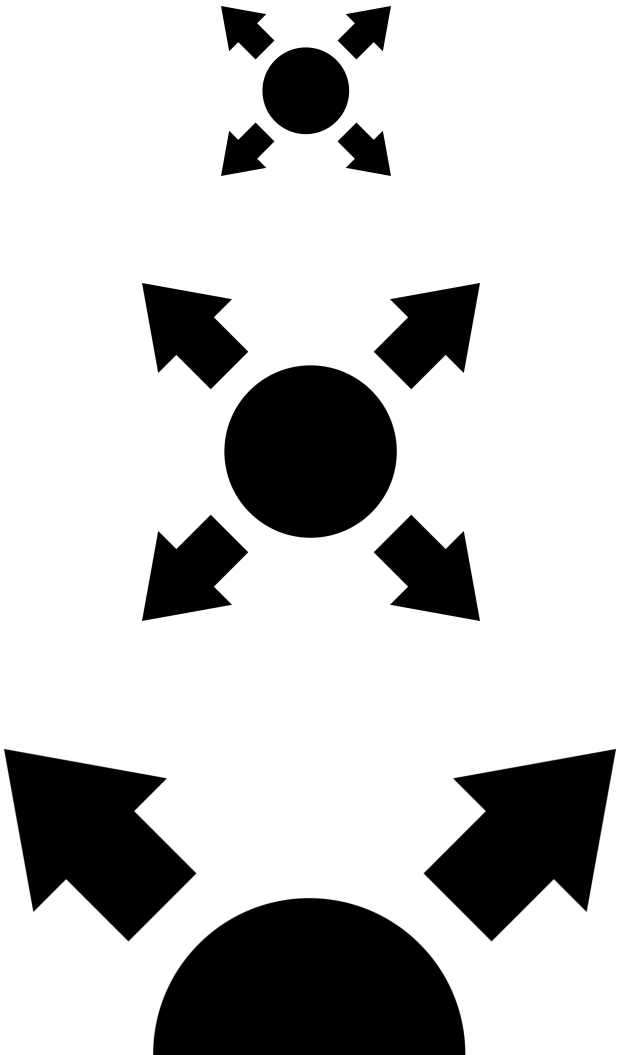


Why Networks?

The document, then, is no longer for history an inert material through which it tries to reconstitute what individuals have done or said, the events of which only the trace remains; history is now trying to define within the documentary materials itself unities, totalities, series, relations.

-Michel Foucault
The Archaeology of Knowledge

Why Networks?



- Scale
  - A few hundred to hundreds of thousands
- See the larger patterns
  - Visualize collection strengths
  - Find outliers

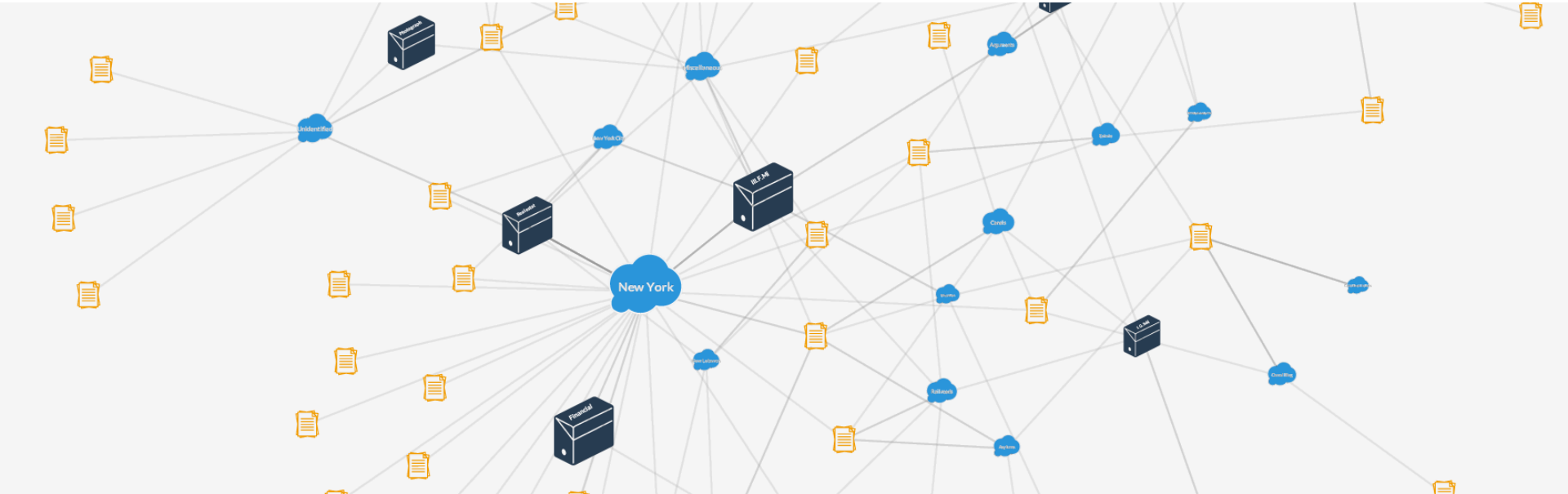
Why Networks?

- Serendipity
  - Interconnect lots of things and you start getting interesting relationships.
  - An alternative way to browse.

Networks in Archives

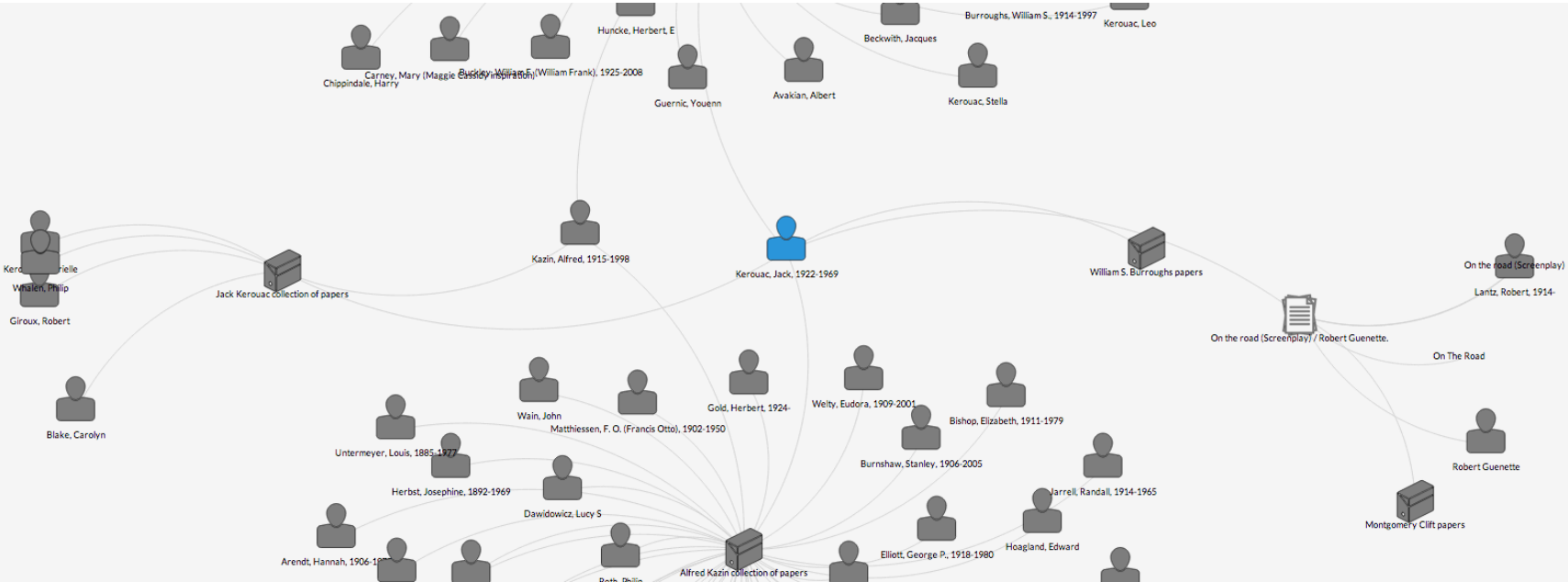
- A single collection viewed as a network.
- http://archives.nypl.org/mss/2993

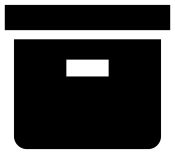
- (shout out to Trevor Thornton, @trevorthornton)



Networks in Archives

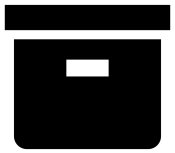
- Explore all of the archive's collection through a network.
- http://archives.nypl.org/terms/





Networks in Archives

- Limits of control access terms.
  - Most collections do not have component level access terms.
  - Collection level terms are not specific to find the interesting connections that we are looking for.
  - We need more data!
- Technical limits of browser based networks.



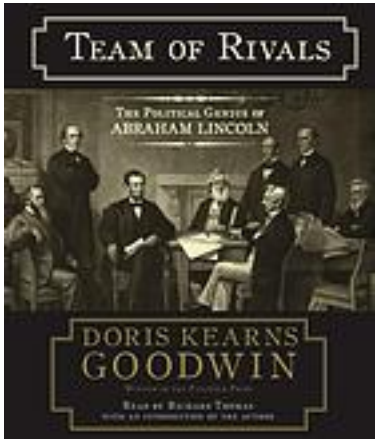
Networks in Archives

- Next steps
  - NLP the EAD files for more controlled terms.
  - Leverage exiting term's linked data to bolster terms associated with a collection.
  - Add institutional data to the graph, who donated what.
  - Start building links between collections based on shared web visits.

Networks in the Catalog

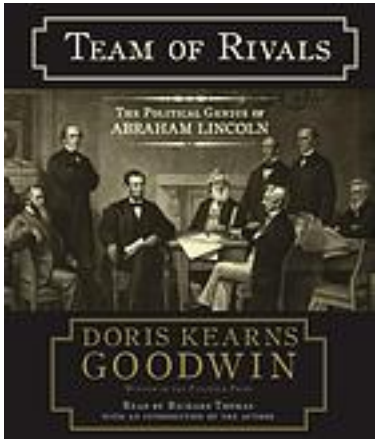
- Each subject is a node in the graph. Including subdivisions.
- A subject is added only if it is used in more than one resource.
- The size of the node is determined by the number of occurrences.
- When two subjects occur in the same record a connection or edge is made between them.
- The more co-occurrences the stronger the connection

Networks in the Catalog



Lincoln, Abraham, -- 1809-1865.
Political leadership -- United States -- History.
Genius -- Case studies.
United States -- History -- Civil War, 1861-1865.
Lincoln, Abraham, -- 1809-1865 -- Friends and associates.
Presidents -- United States -- Biography.
United States -- Politics and government -- 1861-1865.

Networks in the Catalog



Lincoln, Abraham, -- 1809-1865.

Political leadership -- United States -- History.

Genius -- Case studies.

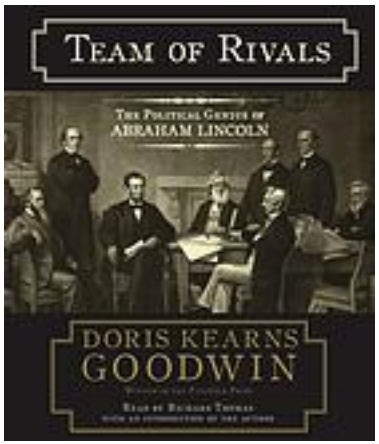
United States -- History -- Civil War, 1861-1865.

Lincoln, Abraham, -- 1809-1865 -- Friends and associates.

Presidents -- United States -- Biography.

United States -- Politics and government -- 1861-1865.

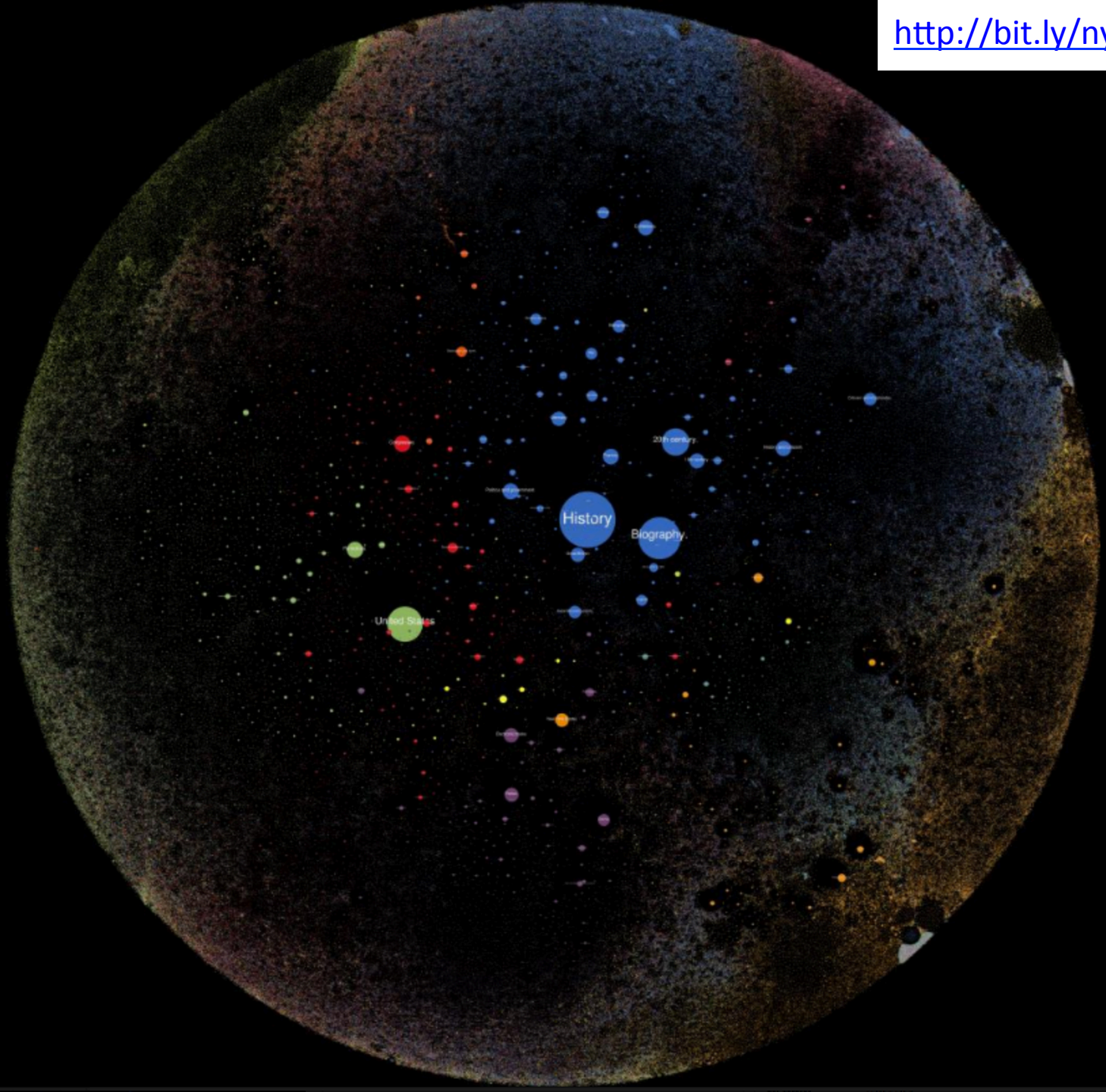
Networks in the Catalog

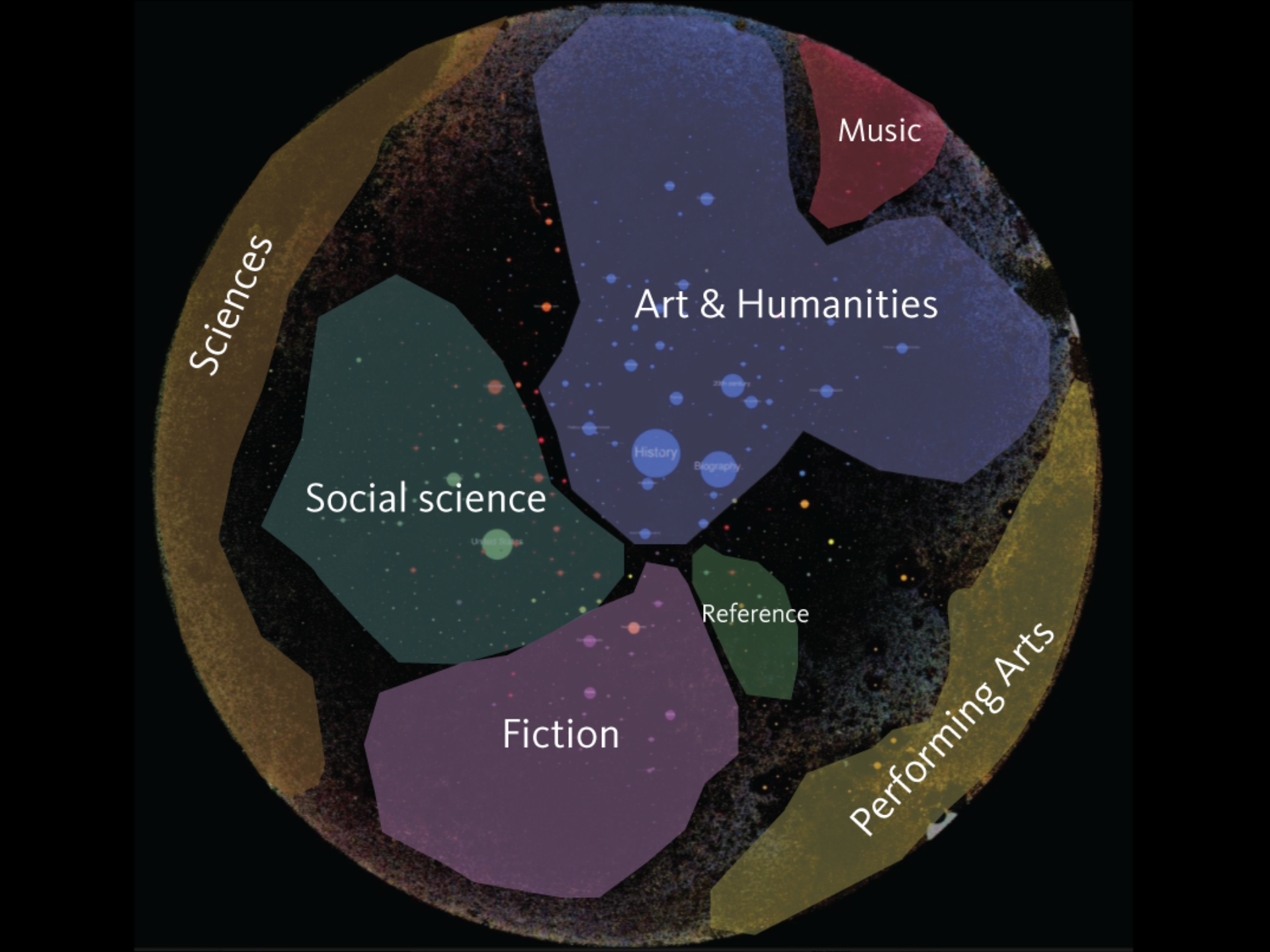


Lincoln, Abraham, -- 1809-1865.
Political leadership -- United States -- History.
Genius -- Case studies.
United States -- History -- Civil War, 1861-1865.
Lincoln, Abraham, -- 1809-1865 -- Friends and associates.
Presidents -- United States -- Biography.
United States -- Politics and government -- 1861-1865.

Networks in the Catalog

- 8.5+ million MARC Records
- 430,000+ subjects
- 11+ million connections


http://bit.ly/nyplnetwork
History
Biography
United States


Music
Sciences
Art & Humanities
History
Biography
Social science
Reference
Fiction
Performing Arts

Networks in the Catalog

- How to:
  - Compile MARC records into node/edge relationships.
  - A tool to render the network based on existing network analysis algorithms.
  - Take that information and make a visual representation of the network.
  - Present that visual representation in a web base format.

Networks in the Catalog

- Compile MARC records into node/edge relationships.
  - pyMARC to process subject headings into a Gephi XML document.

Networks in the Catalog

- A tool to render the network based on existing network analysis algorithms.
  - Using Gephi, but not the GUI.
    - Gephi Toolkit – a library of the core rendering functions of Gephi.
    - https://gephi.org/toolkit/
    - Using the Force Atlas 2 layout algorithm.

- Video. http://vimeo.com/91428315

Networks in the Catalog

- Take that information and make a visual representation of the network.
  - Used Python bindings of GraphicsMagick to draw one very large PNG.
  - Cut into tiles with libvips.
  - Keep track of the coordinates of all nodes.

Networks in the Catalog

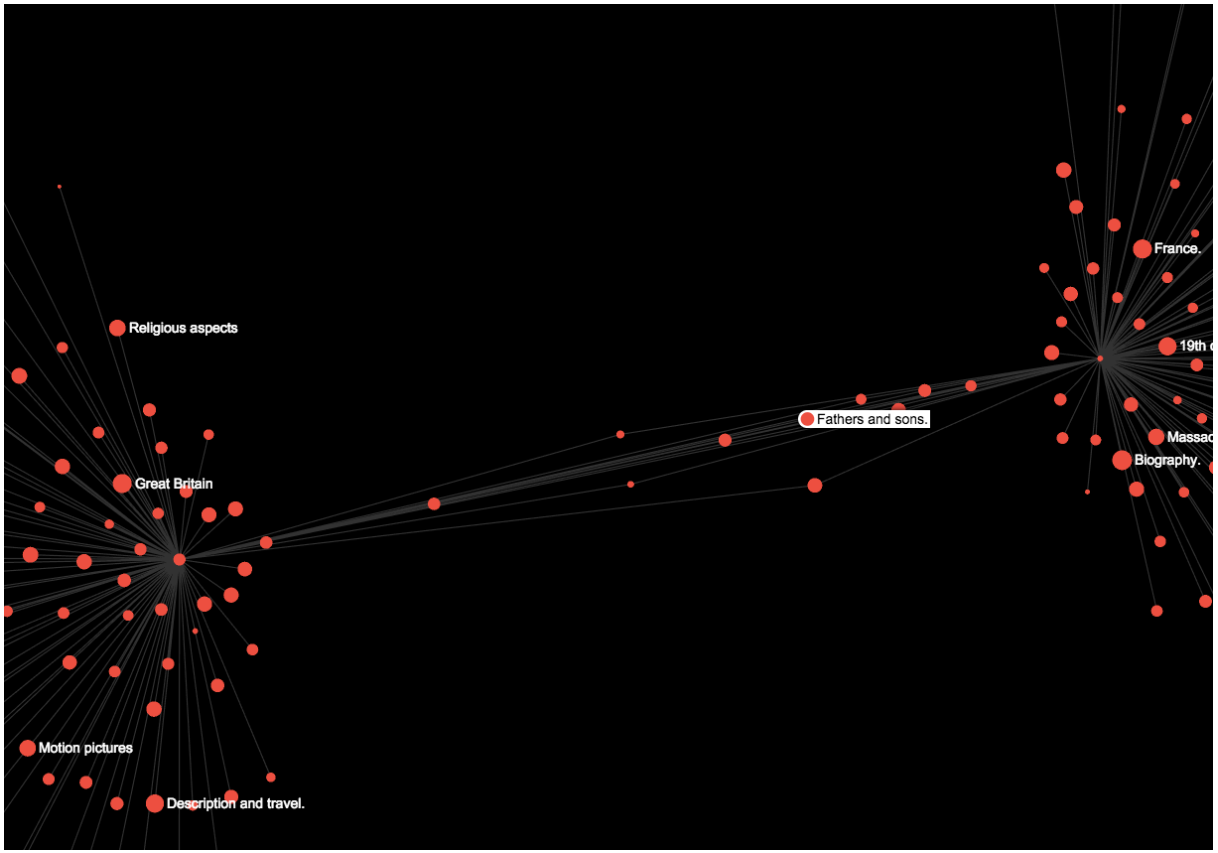
- Present that visual representation in a web base format.
  - Used OpenSeaDragon image tile viewer to navigate.
  - Used ElasticSearch to translate client clicks into node positions and provide search function.

Networks in the Catalog

- Next Steps
  - Continue to refine the rendering parameters and experiment with other layout algorithms.
  - Use subfield subject in more specific ways.
  - Provide more interactivity, sub-networks, shared connections, etc.

Networks in the Catalog

- Next Steps – Bottom up approach



Networks in the Catalog

- Next Steps – What fields to use next?
  - http://bit.ly/nyplcatalogreport
  - What MARC fields can we use next, need to know what we have been populating.

Thanks!

- Matt Miller
  - matthewmiller@nypl.org
  - @thisismmiller
- Tools + Code
  - http://archives.nypl.org/terms
  - http://bit.ly/nyplnetwork
  - https://github.com/thisismattmiller/catalog-network

Icon Credits:
Network designed by Hyeji Michelle Jun from the Noun Project
Book designed by Chris Thoburn from the Noun Project
Box designed by Michael Rowe from the Noun Project
Book designed by Pedro Lalli from the Noun Project
Enlarge designed by Cornelius Danger from the Noun Project
Present designed by Naomi Atkinson from the Noun Project